Plant Virology Protocols, Methods in Molecular Biology.
Ichiro Uyeda and Chikara Masuta (eds.),

Detection and analysis of non-retroviral RNA virus-like elements in plant, fungal and insect genomes

Hideki Kondo*, Sotaro Chiba and Nobuhiro Suzuki

Institute of Plant Science and Resources (IPSR), Okayama University, Kurashiki 710-0046, Japan

*Corresponding author
Hideki Kondo
e-mail: hkondo@rib.okayama.u-ac.jp
Tel./ Fax. +81(86) 434-1232

Word count: summary 114, text 4091
Figures: 5

Abstract

Enodogenous non-retroviral RNA like sequences (NRVSs) have been discovered in the genome of a wide range of eukaryotes. These are considered as fossil RNA viral elements integrated into host genomes by as-yet-known mechanisms, and in many cases, those fossils are estimated to be millions-of-years-old. It is likely that the number of NRVS records will increase rapidly due to the growing availability of whole-genome sequences for many kinds of eukaryotes. Discovery of the novel NRVSs and understanding of their phylogenetic relationship with modern viral relatives provide important information on deep evolutionary history of RNA virus–host interactions. In this chapter, therefore, the common strategies for the identification and characterization of endogenous NRVSs from plants, insects and fungi are described.

Key Words: Pareovirology, Molecular fossil record, Non-retrovirus-like sequence, Database search, Whole-genome shotgun, Genomic PCR, Southern blotting, Phylogenetic analysis, Maximum-likelihood.

1. Introduction

Paleovirology, the study of endogenous viral elements, provides us with important information about the deep evolutionary history of virus–host interactions (1, 2). In eukaryotes, their genomes contain numerous sequences that have originated from retroviruses (reverse-transcribing viruses) whose replication requires integration as proviral DNA into the genome of host cells (2). Endogenous retroviral sequences are the evidence of ancient retroviral infections, thus considered as a kind of molecular fossil record. In contrast, the sequences of non-retroviral RNA viruses, which do not use a DNA intermediate, were until recently considered not to leave such molecular fossils in eukaryotic nuclear genomes. However for the past five years, the rapid progress on whole-genome sequencing for large numbers of eukaryotes has led to the discovery of non-retroviral RNA virus-like sequences (NRVSs, syn. endogenous virus elements: EVEs) integrated into the diverse eukaryotic genomes (1, 2), which are probably the result of heritable horizontal gene transfer (HGT) from viruses to hosts.

The first of these discoveries in vertebrate genomes is a set of NRVSs called EBLN (endogenous bornavirus-like nucleoprotein) which derived from the nucleoprotein gene of an ancient bornavirus (3-5). In addition, NRVSs originated from ancient filoviruses (Ebola and Marburg viruses) have also been found in vertebrate (4-6). Both bornaviruses and filoviruses are negative strand (-)ssRNA viruses belonging to the order *Mononegavirales*. Subsequently, the presence of several NRVSs related to (-)ssRNA viruses in plant, fungal and insect genomes have been reported. NRVSs related to the nucleocapsid protein genes of cytorhabdoviruses and varicosaviruses were found in species of over 9 plant families, including *Brassicaceae* and *Solanaceae* (7). L polymerase-like elements of rhabdoviruses (order

Mononegavirales) and nyaviruses (the proposed members of order *Mononegavirales*) were identified in the genomes of black-legged tick, mosquitoes and several kinds of other insects (4, 8). In fungi, the first (-)RNA virus infection was evidenced based on a discovery of mononegavirus L-like elements in the genome of a phytopathogenic obligate ascomycete, *Erysiphe pisi* (8). For dsRNA viruses, the most widespread NRVSs are related to the capsid protein (CP) and RNA-dependent RNA polymerase genes from partitiviruses and totiviruses in the genomes of plants, arthropods, fungi, nematodes, and protozoa (7, 9). In comparison with dsRNA and (-)ssRNA viruses, there are fewer examples of NRVSs (plant and invertebrate) related to positive-sense (+)ssRNA viruses. These include some members of plant virus genera such as *Benyvirus Cilevirus*, *Citrivirus*, and *Tobamovirus*, and one insect virus genus *Flavivirus* (4, 7, 10, 11).

Here we describe the detail of the methods for the identification and characterization of novel NRVSs from plant, insect and fungal genomes (see Fig. 1). These methods are also applicable to other organisms for NRVSs searches.

2. Materials

2.1. Searching tools for detection of NRVSs in public databases

1. BLAST sequence database search at the National Center for Biotechnology Information (NCBI) (13). A Web server is running at http://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. Phytozome (41 green plant and algal species in version 9.1) (at http://www.phytozome.net). Other local databases for plant genomes: e.g. Brassica database (BRAD) (at http://brassicadb.org/brad/), Sol genomics network (SGN,

Solanaceae Project) (at http://solgenomics.net), Miyakogusa.jp (*Lotus japonicas*) (at http://www.kazusa.or.jp/lotus/blast.html), SoyBase (soybean genetics and genomics database) (at http://soybase.org), MaizeGDB (*Zea mays*) (at http://www.maizegdb.org), Dendrome (a forest tree genome database) (at http://dendrome.ucdavis.edu).

3. FungiDB: an integrated genome database for fungi (46 Fungi and 6 Oomycetes in version 2.3) (at http://fungidb.org/fungidb/) (see Note 1).
4. Arthropod genome databases: e.g. AphidBase (the aphid genome database) (at http://www.aphidbase.com), BeetleBase (*Tribolium castaneum*) (at http://beetlebase.org), FlyBase (a database of *Drosophila* genes and genomes) (at http://flybase.org), Hymenoptera Genome Database (HGD) (at http://hymenopteragenome.org), SilkDB (at http://silkworm.genomics.org.cn), VectorBase (Genomic resources for invertebrate vectors of human pathogens) (at https://www.vectorbase.org).

2.2. Detection and sequencing of NRVSs from plant, fungal and insect materials

1. Materials of interest, i.e., plant seeds or fresh leaf, fungal strains and/or insect individuals.
2. DNA/RNA extraction.
   a. DNA extraction buffer (200 mM Tris-HCl, 250 mM NaCl, 25 mM ethylene diamine tetraacetic acid [EDTA], and 5% [w/v] SDS, pH 7.5).

      Isopropanol.

      1×TE buffer: 10 mM Tris-HCl, 1 mM EDTA, pH 8.0.

   b. DNeasy® Blood and Tissue Kit (Qiagen).

c. 2×CTAB extraction buffer: 2% (w/v) hexadecyltrimethylammonium bromide (CTAB), 100 mM Tris-HCl pH 8.0, EDTA 20 mM pH 8.0, 1.4 M NaCl, 1% (w/v) PVP (polyvinylpyrrolidone, MW 40,000), 0.3% (v/v) β-mercaptoethanol.

d. Phenol (water-saturated for RNA).

Chloroform/isoamyl alcohol (24:1).

Absolute ethanol.

0.1 M sodium acetate, pH 5.

75% Ethanol.

3. Reagents for polymerase chain reaction (PCR), reverse transcription (RT)-PCR, nucleotide sequencing, and Southern blotting.
4. Wizard® SV Gel and PCR Clean-Up System (Promega).
5. Auto Assembler™ DNA Sequence Assembly Software (Applied Biosystems Inc.).
6. GENETYX-MAC/ATSQ (GENETYX Co.) or Enzyme X version 3 (Mek & Tosj) from http://nucleobytes.com/index.php/enzymex.
7. CENSOR, a tool for annotation, submission and screening of repetitive elements in Repbase (at http://www.girinst.org/censor/index.php) (14). CENSOR can be downloaded from the Genetic Information Research Institute (GIRI) for local installation (http://www.girinst.org/censor/download.php) (see Note 2).

2.3. Phylogenetic analyses

1. MAFFT (Multiple alignment program for amino acid or nucleotide sequences) version 7 (at http://mafft.cbrc.jp/alignment/server) (15). MAFFT (version 7.130) is also downloadable from http://mafft.cbrc.jp/alignment/software/.
2. Alignment curing programs.

a. Gap Strip/Squeeze (version 2.1.0), a tool for removing gaps in the alignment, in the Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov/content/sequence/GAPSTREEZE/gap.html).

b. MEGA (Molecular Evolutionary Genetics Analysis) version 4.02 software (from http://www.megasoftware.net/mega4/mega.html) (16). A current version for MEGA (version 6 for windows or version 5 for Mac OS) is downloadable from http://www.megasoftware.net.

3. ProtTest server (current version is 2.4), a bioinformatics tool for the selection of best-fit models of protein evolution (at http://darwin.uvigo.es/software/prottest_server.html) (17). ProtTest (version 3.2.1) is downloadable from http://code.google.com/p/prottest3/.
4. PhyML 3.0, a tool for estimating maximum-likelihood (ML) phylogenies (at http://www.atgc-montpellier.fr/phyml/) (18). A new release of PhyML (version 3.1) is also available at http://www.atgc-montpellier.fr/phyml/versions.php.
5. FigTree version 1.3.1 software, a tool for drawing the tree (from http://tree.bio.ed.ac.uk/software/).

3. Methods

Here we describe a protocol for NRVS searches in section 3.1., a brief procedure for confirmation experiments in section 3.2., and a method of maximum-likelihood phylogenetic analyses of NRVSs in section 3.3., which were basically used in our previous studies (see Fig. 1). Examples of results obtained in the above processes (based on discovery of NRVSs) are also provided.

3.1. Genome sequence database search

1. To screen host genomic sequences for NRVSs, prepare query viral sequences. We usually select type species of the genera (non-retroviral RNA viruses) and obtain viral sequences from the NCBI taxonomy browser (http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=Viruses). Queries should be amino acid sequences of the entire or conserved domain region of viral proteins (see Note 3). Those of viruses of your interest would be also preferable.
2. Conduct BLAST (tBLASTn) searches with prepared query sequences against genome sequence databases available from the NCBI (nucleotide collection, nr/nt; genome survey sequences, GSS; high-throughput genomic sequences, HTGS; whole-genome shotgun contigs, WGS, and others) with the default parameter setting.
3. As an option, other local databases would be used for NRVS searches. Specific sequence data warehouses such as Phytozome, FungiDB and other arthropod genome databases usually include both updated and newly released genomes (see Note 1).
4. Genome sequences that matched viral peptides with E-values smaller than $1e^{-5}$ (the most common conventional value) are extracted as candidate NRVSs together with their flanking sequences if available (see Note 4).
5. Analyze candidate sequences and determine the region covering possible viral elements. GSS-, HTGS- and WGS-derived sequences often provide only partial fragments, but in many cases partially or entirely overlapping fragments can be found in a series of search. Thus it is possible to extend given NRVSs by assembling of those.

6. If NRVSs have interrupted ORFs, restore them by adding single or double “N” at frame-shifting sites or by introducing triplet “N” instead of internal stop-codons where they can be inferred by tBLASTn alignment (see Fig. 2A and Note 5). This step is required to obtain continuous, deduced amino acid sequences of NRVSs for further analyses. Edited residues are shown as “Xs” and these would be taken into account for mutations occurred during the course of evolution.
7. Each extracted candidate is then used as a reverse tBLASTn query against the non-redundant (nr) database (see Fig. 2A). This step is helpful to detect related NRVSs as well as more evolutionally close viruses than query viruses. These NRVSs are also use to screen WGSs of other organisms, ESTs (non-human, non-mouse expressed sequence tags) and TSAs (transcriptome shotgun assembly) available from the NCBI (see Note 6).
8. To examine for the presence of potential transposable elements and repetitive sequences in NRVSs and their flanking sequences, candidates (non-restored sequences) are subjected to the CENSOR program provided by the GIRI (see Fig. 2A and Note 7).
9. Nomenclature of NRVSs. Currently no fixed rule for nomenclature of fossil viral elements is present, and you may find many terms depending on reports. A system we use is, for example, as follows: “host organism name” + “viral protein” + “-like sequence” + “numbers defining virus species”, i. e., *Arabidopsis thaliana* partitivirus CP-like sequence 1 (AtPCLS1), *Erysiphe pisi* mononegavirus L protein-like sequence 1 (EpMLLS1), etc (see legend for Figs. 2, 4 and 5).
10. Deposition of NRVSs with appropriate annotations in GenBank/EMBL/DDBJ would be helpful for further studies. Nevertheless, this should be restricted to ones determined here by genomic PCR and sequencing (see below).

3.2. Detection and sequencing of NRVSs from Plant, fungal and insect materials

1. To experimentally confirm the presence of NRVSs in the plant, fungus and insect chromosomes, obtain each material of interest (i.e., plant seeds, fungal strains or insect individuals) from commercially available materials, public stock centers, laboratories and/or fields.
2. DNA/RNA preparation (in our system).
   a. For plant genomic DNA, seeds or fresh leaf materials are homogenized in a micocentrifuge tube containing DNA extraction buffer (19). Homogenates are subsequently centrifuged at 15,000 *g* for 10 min. Nucleic acids in the supernatant are then precipitated by centrifugation as above after addition of 0.6 volume of isopropanol. Resuspend the pellet in TE buffer and use as a template for genomic PCR (see Note 8).
   b. For fungal and insect materials, total genomic DNA are purified using the DNeasy® Blood and Tissue Kit (Qiagen) according to the manufacturer's instructions (see Note 8).
   c. The genomic DNA isolation for Southern blot analysis is performed following a standard CTAB protocol (20).
   d. Total RNA fractions are obtained from fresh leaf materials of interest by two rounds of phenol-chloroform extraction and one round of chloroform extraction, followed by ethanol precipitation.
3. To amplify the NRVS fragments from DNA samples by PCR, specific primer pairs are designed based on the virus-related sequences and their flanking sequences (see

Fig. 2B and Note 9). As a control, primer sets to amplify a fragment of the ribosomal internal transcribed spacer (ITS) region are also used (see Note 10).

4. PCR is generally carried out in a final volume of 50 µl containing 25 µl of Quick Taq HS DyeMix (TOYOBO) with the following conditions: an initial 94˚C for 2 min, followed by 35 cycles of 94˚C for 10 sec, 58˚C for 30 sec, and 68˚C for 1 min, then 68 ˚C for 10 min (see Note 11).
5. Subject the PCR product to 1% agarose gel electrophoresis. Stain the gel with ethidium bromide and visualize DNA bands under UV illumination.
6. Sequence the purified PCR fragment. We use an ABI3100 DNA sequencer (Applied Biosystems) with BigDye® Terminator chemistry and a specific primer (see Note 12).
7. Assemble resultant sequences into a single fragment. We use the Auto Assembler software for the sequence assembly and analyze the sequences using the DNA processing software packages (GENETYX or Enzyme X). Finally, compare actual NRVSs and counterparts from databases, and confirm the presence of NRVSs on the genome of tested organisms.
8. (Option 1) To know the copy number of the NRVSs in chromosomes, we recommend performing Southern blot analysis as described by Faruk et al. (21). Digoxigenin (DIG)-11-dUTP-labeled DNA fragments are amplified as probes from genomic DNA according to methods recommended by the manufacturer (Roche Diagnostics).
9. (Option 2) To test whether NRVSs could be expressed in cells, the BLASTx and BLASTp searches against databases for ESTs and TSAs available from the NCBI are a relatively easy way. In addition, RT-PCR detection of NRVS transcripts using an RNA template is more preferable (see Note 13). A separate reverse-transcription (M-MLV, Invitrogen) and PCR (see step 4) reactions are conducted with specific

primer pairs for each NRVS (see Fig. 3). One-step RT-PCR kits (e.g., OneStep RT-PCR Kit, Qiagen) are also recommended to provide a fast and successful alternative.

3.3. Phylogenetic analyses

1. The deduced amino acid sequences of NRVSs (in a restored form) are used for the phylogenetic analysis with related protein sequences from extant viruses. Prepare a sequence list of those in FASTA format.
2. Multiple alignments of amino acid sequences are constructed using MAFFT version 7 (see Note 14). Copy and paste the prepared sequence list to the window of the program, then run under the default parameters.
3. Reformat the resultant alignment with "ReadSeq" (copyright 1990 by D. G. Gilbert) in the same site to convert between the different sequence file formats used by following programs.
4. Automatic sequence alignments generally contain numerous gaps, therefore we recommend to refine the alignment by removing gap regions (columns) (see Note 15).
   a. Gap columns in the multiple alignments can be removed by using the Gap Strip/Squeeze program online.
   b. The alignment can be manually edited in MEGA software.
5. To obtain appropriate substitution models for the maximum likelihood (ML) analyses, each of cured data sets (PHYLIP formatted alignment) should be subjected to the "AIC" (Akaike information criterion) calculation using ProtTest server (an example of output, AIC: LG+I+G+F). You may receive an e-mail notification when ProtTest

analysis is done.

6. A phylogenetic tree is generated using the appropriate substitution model (see below) in PhyML 3.0 at the ATGC bioinformatics platform.

   a. Input above cured data (PHYLIP formatted alignment).

   b. Select the appropriate substitution model and other specific improvements determined above; “F” (amino acid frequencies; optimized or empirical), “I” (proportion of invariable sites; fixed or estimated) and “G” (gamma distribution parameter; fixed or estimated). PhyML 3.0 uses the “LG” substitution model as default.

   c. The tree searching algorithms (the type of tree improvement) provided by PhyML are “NNI” (nearest neighbor interchange, a fast algorithm) and “SPR” (subtree pruning and regraft, a slower but efficient algorithm) (22). The best option here is probably to use a SPR search (see PhyML-Manual version 3, available http://www.atgc-montpellier.fr/phyml/usersguide.php).

   d. Selection of the method used to measure branch support. PhyML 3.0 provides “aLRT” (approximate likelihood ratio test, a fast algorithm) for convincing the branching accuracy, which is a good alternative to the bootstrap analysis (time-consuming). The default is to use “SH-like” (Shimodaira–Hasegawa–like) procedure (23) (see Note 16).

   e. PhyML generates a tree file and a model parameter file. The estimated maximum likelihood tree is in the standard “Newick” format. You may receive an e-mail with a compressed Zip file containing these materials.

8. Open the tree file (Newick format) in the FigTree program. Select “Midpoint Rooting” from the view menu or “Reroot” button to do outgroup rooting if you include appropriate outgroup(s). The node/branches of the tree (aLRT-SH like values) can be

labeled from the “Display” in the Node Labels menu. The tree can be exported as a graphic file by selecting “Export Graphic” from the file menu. We are using the Mac OSX version here, but there is also a Windows version. Obtained trees are modified in the illustrator software (or in any drawing tools) by keeping node length and scale bar. See Figs. 4 and 5 for the phylogenetic analysis of NRVSs.

9. (Option 3) Timing of NRVS-integrations into host chromosomes is of particular interest. This requires both information on divergence timing of host species within given taxa and on substantial variety of the NRVS locating on the same locus of chromosomes in related organisms. Based on detection profile of the NRVS in related organism, the integration timing can be estimated from host-divergence timing; the integration should have occurred before branching of NRVS-containing groups but after branching of those with outer groups which lack the NRVS in the same locus. See an example of AtPCLS1 reported by Chiba et al. (6) (see also Note 17).

4. Notes

1. Searching local databases for the fossil records may also be important for the discovery of novel NRVSs. In our previous study, we identified (-)ssRNA virus like sequences (NRVSs) in the WGS assemblies of the chromosomes of a plant pathogenic obligate ascomycete (powdery mildew fungus), *Golovinomyces orontii* (order Erysiphales) (8) (see Fig. 5), which are available from the Max Planck Institute for Plant Breeding Research Powdery Mildew Genome Project site (http://www.mpipz.mpg.de/24322/Project_Description), but not from the NCBI.

2. RepeatMasker may also useful to identify, characterize, and analyze repetitive elements in genomic sequences. The program can be run on a web server at http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker.
3. Longer and/or multiple queries for WGS searching (e.g. entire amino-acid sequences for viral polyprotein or replicase) tend to be aborted with the message "Error: CPU usage limit was exceeded." Therefore, we recommend the shorter queries (e.g., conserved domain regions for viral proteins) and/or selecting against the specific plants (taxid:3193), fungi (taxid:4751) or insects (taxid:6960) by using "Choose Search Set" to search for novel NRVSs through tBLASTn.
4. In our previous papers, we use <0.01 as the cut-off value for a "match", because some hits with values smaller than 0.01 show higher values in the reverse BLAST analyses against identified NRVSs (7, 11). In addition, although some viral genes had sequence similarity with plant helicase or heat shock proteins, we removed those from further considerations, since these are likely originated from horizontal gene transfer events form host organisms to viruses (7).
5. Although several NRVSs in plant chromosomes retain "in-frame" ORFs (7), mononegavirus-like sequences in fungal chromosome and benyvirus-like fragments in plant and insect chromosomes are mostly fragmented (8, 11).
6. Several studies have demonstrated that the transcriptome analysis based on next-generation sequencing technologies is a useful new research tool for the discovery of RNA viruses. Thus, this reverse BLAST searching has another important aspect for discovering potential novel viruses. In fact, during BLAST searching for (-)ssRNA virus like sequences, we found evidence strongly suggesting the presence of extant (-)ssRNA viruses in the kingdom Fungi for the first time (8).

7. Many NRVSs have distinguishable flanking sequences of the host origin in which these contain trace putative transposable elements (7, 8, 11) (see Fig. 2). From these data, it is speculated that progenitor viral segments might have been integrated after reverse-transcription with the aid of retrotransposons.

8. Purified DNA solutions (TE buffer) were stored at 4 °C until use.

9. This step importantly confirms whether candidate NRVSs are of chromosomal sequences or contaminating sequences from exogenous viruses.

10. For plant genome, a primer set, At-IRS-FW (ITS-F: CCGTAGGTGAACCTCGGAGGG) and At-IRS-RV (ITS-R: GGTGATCCCGCCTGACCTGG) (7), are used for amplification of the ITS regions 1 and 2 including the 5.8S rDNA.

11. PCR conditions should be optimized for each DNA template.

12. PCR fragments may have a few varieties in their sequences if NRVSs are multiplied after integration. To analyze multiple copies of NRVSs, PCR fragments from these loci should be cloned in pGEM T-easy (Promega) or other suitable vectors and then sequenced.

13. Some NRVSs having a long “in-frame” ORF appear to be transcribed (see Fig. 3B) and they might be translated as cellular proteins. However, further studies are required to determine their molecular function in the cells. It should be noted that the CP of a fungal partitivirus (Rosellinia necatrix partitivirus 2) had the greatest sequence similarity to a plant gene product, *Arabidopsis thaliana* auxin indole-3-acetic acid (IAA) Leucine resistant 2 (ILR2), which is thought to regulate a part of plant hormone homeostasis (7, 9).

14. It is also useful for refining multiple alignments obtained by other methods, e.g. T-Coffee (from http://www.tcoffee.org) and ClustalW (from http://www.clustal.org/clustal2/) programs.
15. We also recommend the removal (masking) of the align parts with low confidence (potentially misaligned) using the Gblocks, a well known program to remove poorly aligned regions (from http://molevol.cmima.csic.es/castresana/Gblocks.html) (24). However, in our previous studies, some alignments appeared to be too short for the elimination of problematic regions by Gblocks. Thus we used only Gap Strip/Squeeze or MEGA program.
16. The sets of branching support values with bootstrap proportion >0.75 and aLRT>0.9 (SH-like option) tend to be similar (see an user guide at http://www.atgc-montpellier.fr/phyml/usersguide.php?type=online).
17. Using endogenous viral elements, bornavirus fossils were determined to be more than 40 million years old in primates, and even much older in other mammalians (3). Endogenous filovirus (Ebola and Marburg viruses) fossils were also identified and estimated to be at least 20 million years old in rodents and 40 million years old in other mammalians (5).

Acknowledgment This work was supported in part by a Grant-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology (KAKENHI 24580064) (H.K. and N.S.), and the Sanyo Broadcasting Foundation for Science and Culture (HK).

References

1. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. Nature Rev Genet 13:283–288.
2. Patel MR, Emerman M, Malik HS (2011) Paleovirology - ghosts and gifts of viruses past. Curr Opin Virol 1:304–309.
3. Horie M, Honda T, Suzuki Y, et al (2010) Endogenous non-retroviral RNA virus elements in mammalian genomes. Nature 463(7277):84–90.
4. Katzourakis A, Gifford RJ (2010) Endogenous viral elements in animal genomes. Plos Genet 6:e1001191.
5. Belyi VA, Levine AJ, Skalka AM (2010) Unexpected inheritance: Multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes. Plos Pathog 6:e1001030.
6. Taylor DJ, Leach RW, Bruenn J (2010) Filoviruses are ancient and integrated into mammalian genomes. BMC Evol Biol 10:193.
7. Chiba S, Kondo H, Tani A, et al (2011) Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. Plos Pathog 7:e1002146..
8. Kondo H, Chiba S, Toyoda K, Suzuki N (2013) Evidence for negative-strand RNA virus infection in fungi. Virology 435:201–209.
9. Liu HQ, Fu Y, Jiang D, et al (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol 84:11876–11887.
10. Cui J, Holmes EC (2012) Endogenous RNA viruses of plants in insect genomes. Virology 427:77–79.

11. Kondo H, Hirano S, Chiba S, et al (2013) Characterization of burdock mottle virus, a novel member of the genus *Benyvirus*, and the identification of benyvirus-related sequences in the plant and insect genomes. Virus Res 177:75–86.

12. Taylor DJ, Bruenn J (2009) The evolution of novel fungal genes from non-retroviral RNA viruses. BMC Biol 7:88.

13. Altschul SF, Madden TL, Schäffer AA, et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

14. Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

15. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinf 9:286–298.

16. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599.

17. Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

18. Guindon S, Dufayard JF, Lefort V, et al (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321.

19. Miura E, Kato Y, Matsushima R, et al (2007) The balance between protein synthesis and degradation in chloroplasts determines leaf variegation in *Arabidopsis* yellow variegated mutants. Plant Cell 19:1313–1328.

20. Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA

extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Rep 15:8–15.

21. Faruk MI, Eusebio-Cope A, Suzuki N (2008) A host factor involved in hypovirus symptom expression in the chestnut blight fungus, *Cryphonectria parasitica*. J Virol 82:740–754.

22. Hordijk W, Gascuel O (2005) Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. Bioinformatics 21:4338–4347.

23. Anisimova M, Gascuel O (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol 55:539–552.

24. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–577.

25. Kondo H, Kanematsu S, Suzuki N (2013) Viruses of the white root rot fungus, *Rosellinia necatrix*. Adv Virus Res 86:177–214.

Figures

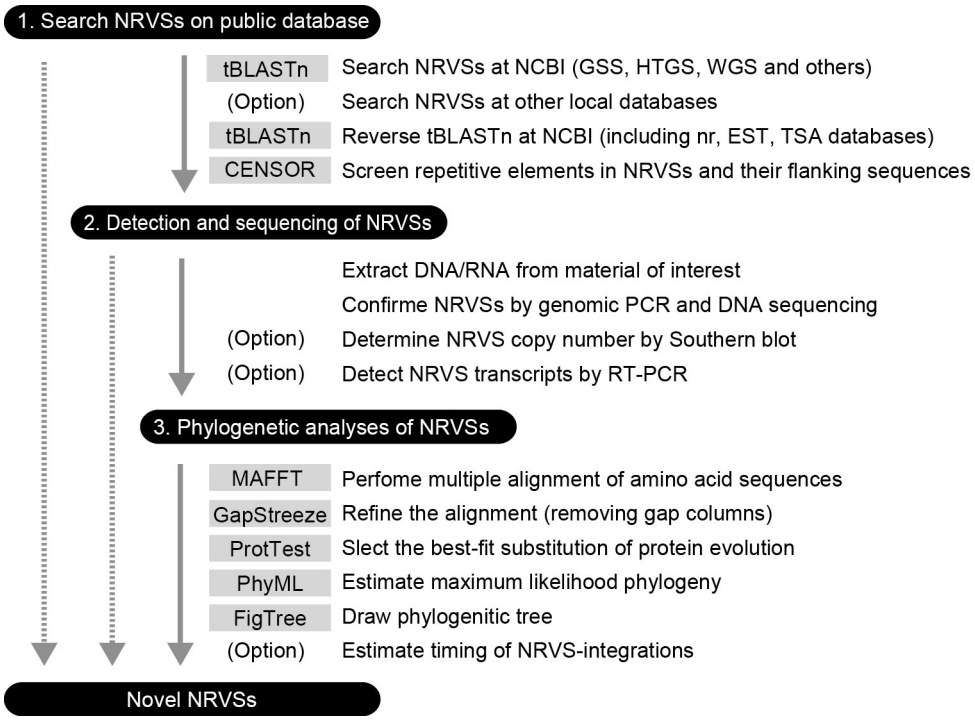

Fig. 1. Scheme for the detection and analysis of endogenous non-retroviral RNA virus-like sequences (NRVSs).

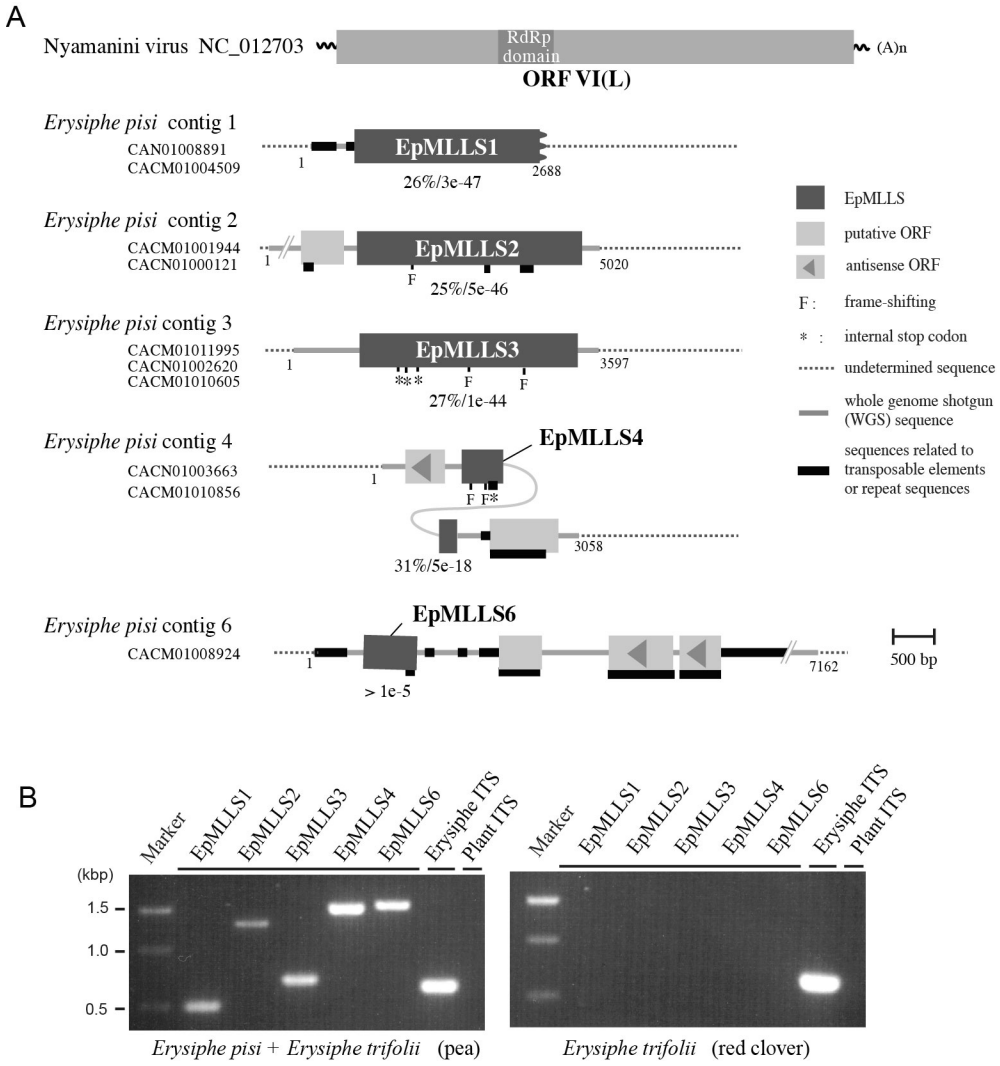

Fig. 2. Negative-strand RNA virus-related sequences from fungal nuclear genome. A. Schematic representation of selected *Erysiphe pisi* (mononegavirus L protein-like sequences) MLLSs and their flanking regions. The corresponding positions of EpMLLSs on the Nyamanini virus (a prototypic member of *Nyavirus*) L-polymerase mRNA. The potential coding regions of EpMLLSs and flanking small ORFs are shown as boxes. EpMLLSs are distantly related to Nyamanini viruses (tBLAST % identities and e-value are shown). Retrotransposon-like sequences are shown by thick black lines. B. Genomic PCR analysis of EpMLLSs. EpMLLSs of the *Erysiphe* spp. field samples

isolated from pea (right panel) and red clover (left panel) were amplified using a primer set specific for each EpMLLS. Primer pairs, EryF (TACAGAGTGCGAGGCTCAGTCG) and EryR (GGTCAACCTGTGATC CATGTGACTGG) and At-IRS-FW and At-IRS-RV, were used for amplification of the *Erysiphe* spp. and plant ITS regions, respectively. This analysis was in part reported by Kondo et al. (8).

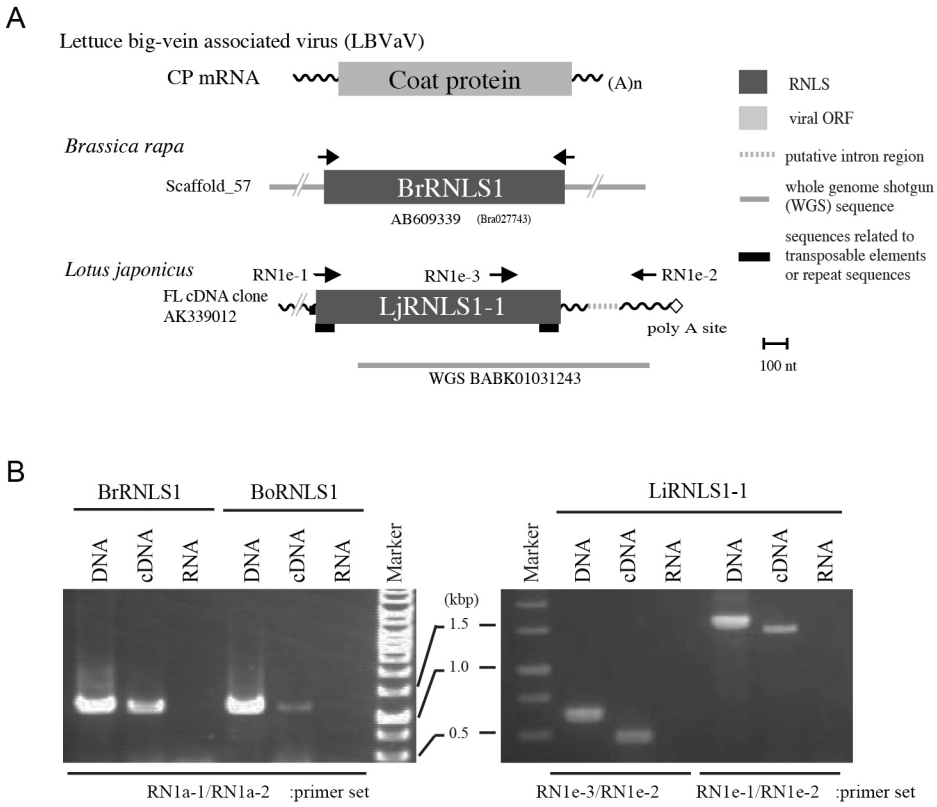

Fig. 3. Expression of negative-strand RNA virus-related sequences from plant nuclear genomes. A. Schematic representation of selected rhabdovirus N-like sequences (RNLSs). RNLSs found in the genome sequence database of *Brassica rapa* (BrRNLS1) and *Lotus japonicus* (LjRNLS1-1) have significant sequence similarity to CP from lettuce big-vein associated virus (LBVaV, Varicosavirus) (7). Note that varicosaviruses are evolutionally related to the family *Rhabdoviridae*. The positions of the primers used for genomic PCR are shown by arrows. B. Molecular detection of RNLSs from *Brasica* and *Lotus* plants. Representative RNLSs from *B. rapa* (BrRNLS1), *B. oleracea* (BoRNLS1) and *L. japonicus* (LjRNLS1-1) were detected by genomic PCR (DNA) as well as by RT-PCR (cDNA) but not by RNA-temperate PCR (RNA). The size differences between genomic PCR (lanes DNA) and RT-PCR (lanes cDNA) products in LjRNLS1-1 verified the presence of an intron region (dot-line in A).

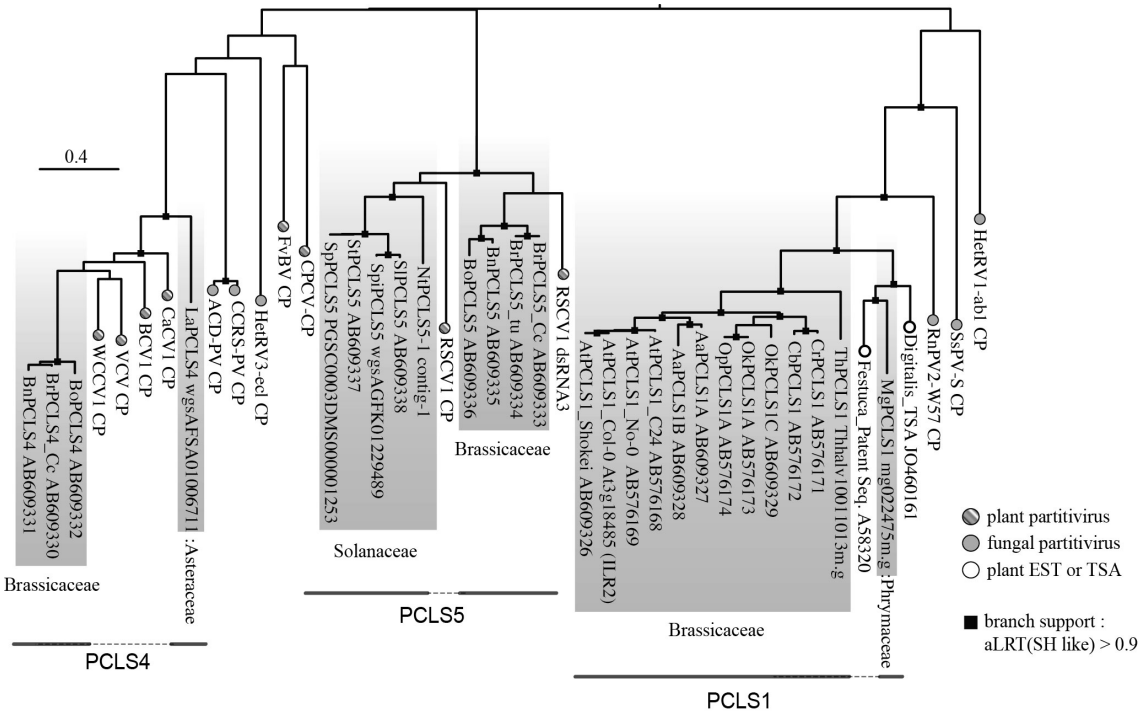

Fig. 4. Phylogenetic tree of selected partitivirus coat proteins (CPs) and partitivirus CP-like sequences (PCLSs) present on plant genomes. An alignment of CP sequences of representative partitiviruses and PCLSs was analyzed by the ML method. The accession numbers are shown next to the sequence names in the figure. Gray-shaded sequences represent NRVSs (PCLSs). A part of this analysis was reported by Chiba et al. (7) and Kondo et al. (25). Plant PCLS1: AtPCLS1 (*Arabidopsis thaliana*), AlPCLS1 (*A. lyrata*), AaPCLS1 (*Arabidopsis arenosa*), OkPCLS1 (*Olimarabidopsis korshinskyi*), OpPCLS1 (*O. pumila*), CrPCLS1 (*Capsella rubella*), CbPCLS1 (*C. bursa-pastoris*), ThPCLS1 (*Thellungiella halophila*), MgPCLS1 (*Mimulus guttatus*); PCLS4: BrPCLS4 (*Brassica rapa*), BoPCLS4 (*B. oleracea*), BnPCLS4 (*B. napus*), LaPCLS4 (*Lactuca sativa*); PCLS5: BrPCLS5 (*B. rapa*), BoPCLS5 (*B. oleracea*), BnPCLS5 (*B. napus*);

StPCLS5 (*Solanum tuberosum*), SpPCLS5 (*S. phureja*), SpiPCLS5 (*S. pimpinellifolium*), NtPCLS5 (*Nicotiana tabacum*). Best model according to AIC: WAG+I+G+F. Closed boxes on the nodes represent aLRT values derived using an SH-like calculation (only values greater than 0.9 are shown).

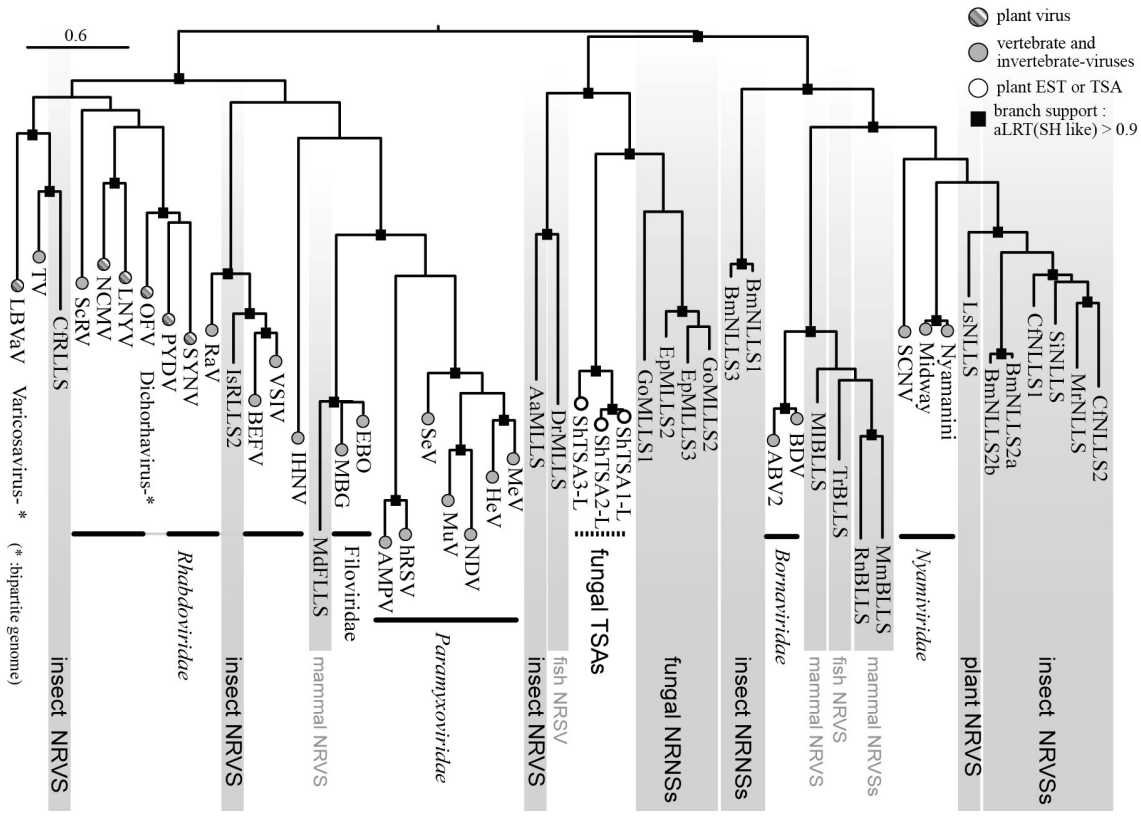

Fig 5. Phylogenetic relationship of L protein sequences of mononegaviruses (filoviruses, paramyxoviruses, rhabdoviruses and bornaviruses) and endogenous mononegavirus L protein-like sequences (MLLSs) from fungi, plants, insects, fish and mammals. This ML-tree was constructed using PhyML 3.0 based on a multiple amino acid sequence alignment of the RdRp polymerase core module. Gray-shaded sequences represent NRVSs (MLLSs) (8). Fungal MLLS: EpMLLS2, 3 (a pea powdery mildew fungus, *Erysiphe pisi*), GoMLLS1, 2, (another powdery mildew fungus, *Golovinomyces orontii*); Insect MLLS: AaMLLS (the yellow fever mosquito, *Aedes aegypti*); Plant NLLS (nyavirus L protein-like sequence, one form of MLLSs): LsNLLS (lettuce, *Lactuca sativa*); insect NLLSs: BmNLLS1, (silkworm, *Bombyx mori*), MrNLLS, (leafcutter bee, *Megachile rotundata*), CfNLLS (carpenter ant, *Camponotus floridanus*),

SiNLLS (fire ant, *Solenopsis invicta*); insect RLLSs (rhabdovirus L protein-like sequence): CfRLLS (*C. floridanus*), IsRLLS (black-legged tick, *Ixodes scapularis*); host names for other MLLSs from mammal and fish genomes are not shown. Fungal TSA-derived L-like sequences: ShTSAs (*Sclerotinia homoeocarpa* transcriptome shotgun assembly). Best model according to AIC: LG+G+F. Closed boxes on the nodes represent aLRT values derived using an SH-like calculation (only values greater than 0.9 are shown).